Combining Social and Intrinsically-Motivated Learning for Multi-Task Robot Skill Acquisition

Thibaut Kulak
Idiap Research Institute
Email: thibaut.kulak@idiap.ch

Sylvain Calinon
Idiap Research Institute
Email: sylvain.calinon@idiap.ch

***Abstract*—This article proposes an approach for coupling internally-guided learning and social interaction in the context of a multi-task robot skill acquisition framework. More specifically, we focus on learning a parametrized distribution of robot movement primitives by combining active intrinsically-motivated learning and active imitation learning. We focus on the case where the learning modalities to use are not specified in advance by the experimenter, but are chosen actively by the robot through experiences. Such approach aims at combining experiential and observational learning as efficiently as possible, by relying on a skill acquisition mechanism in which the agent/robot can orchestrate different learning strategies in an iterative manner, and modulate the use of these modalities based on previous experiences. We demonstrate the effectiveness of our approach on a waste throwing task with a simulated 7-DoF Franka Emika robot, where at each iteration of the learning process the robot can actively choose between observational/imitation learning and experiential/intrinsically-motivated learning.**

I. Introduction

Humans and other animals acquire and refine skills in an open-ended manner through lifelong learning, and are hence autonomous and versatile for interacting and learning in their environments. Despite the important progress in Artificial Intelligence, robots still lack this capacity. Endowing robots with the capability to autonomously discover and solve multiple tasks incrementally and in an open-ended manner is one of the greatest challenges of robotics today and is the goal of the growing field of developmental robotics [1]. In particular, humans have the ability to use several learning modalities, and most interestingly to arbitrate their choice based on their reliability [2]–[4]. In this article, we explore a possible route towards such a goal by proposing a principled computational approach combining intrinsically-motivated learning and imitation learning.

In robotics, skills acquisition is most often studied by concentrating on a single learning strategy, or by predefining a basic sequence of learning strategies in advance (e.g., a reinforcement learning problem initialized with a demonstration). This led to large research efforts dedicated to the development of very elaborated algorithms specialized in a single domain (learning from demonstration, reinforcement learning, curiosity-driven learning). We argue that this complexity could be reduced by allowing several learning strategies, and by providing a mechanism to select these learning modalities in an open-ended and interactive manner. In the same way as we cannot learn to play football only by watching TV and that we cannot learn football tactics from scratch only based on the rules of the game, we believe that robots should rely on multiple learning strategies, whose sequence can only be determined during the course of learning, in a lifelong learning fashion.

The above argument is motivated by studies in various fields including cognitive science [1, 5], ethology [3, 4], neurocomputing [2, 6] and robotics [7]–[11], all proving insights, in different forms, about the importance of combining multiple learning modalities to acquire skills. In particular, several developmental studies such as [3, 4, 12] have shown that learning by imitation is a key component of social learning in child development. Children tend to imitate what they are shown, even if some of the observed actions are not necessarily useful.

From a developmental robotics point of view, we argue that orchestrating multiple learning strategies during the skill acquisition process can better cope with the specific advantages and limitations of each individual strategy. Indeed, these strategies are often complementary to each other, hence the necessity to combine them. Intrinsically-motivated learning requires no external guidance, i.e., no presence of a human, but it usually involves a long interaction process with the environment. Imitation learning, on the other hand, requires the presence of a human, but demonstrations provide a lot of information which would have required a tremendous amount of time to autonomously acquire.

In this article, we propose an active learning approach that can act on different fronts: at a meta-level, by deciding about the currently most appropriate learning modality in an open-ended manner, and at a low-level, by deciding about which of the condition/situation/context the agent currently needs to experience on its own or request as demonstration.

Our contribution is a Bayesian computational framework for learning robot movement primitives providing this high-level and low-level arbitration capability, namely:

- Strategy selection: the robot chooses actively between imitation learning and intrinsically-motivated learning, based on its previous experiences.
- Demonstration choice: in the imitation learning strategy, the robot chooses actively the goal that is expected to yield the most interesting demonstration.
- Policy exploration: in the intrinsically-motivated learning strategy, the robot chooses actively which movement is

going to improve the most its knowledge of the task.

To the best of our knowledge, our work is the first to integrate these three learning aspects in a computational framework.

Our paper is organized as follows. First, we review the existing literature in Sec. II. In Sec. III, we introduce our Bayesian computational framework, and in Sec. IV, we derive two active learning strategies as well as an arbitration strategy. Our experimental results are presented in Sec. V.

II. Related Work

A. Intrinsically-motivated learning

Intrinsically-motivated learning (a.k.a. curiosity-driven learning) has emerged as an efficient approach for autonomous lifelong learning in robots [13, 14]. It is inspired by the ability of humans to discover how to produce interesting effects in their environments [15]–[17]. In [17], psychologists suggested that exploration might be triggered and rewarded for situations that include novelty/surprise. They observed that the most rewarding situations were those with an intermediate level of novelty, between already familiar and completely new situations. This also seems to be confirmed by recent neuroscience studies showing that dopamine might be released, not only for predicting external rewards such as food, but also for internal rewards such as prediction errors [18]. This suggests that intrinsic motivation systems might be present in the brain, potentially by the presence of signals related to prediction errors. Given this background, a way to implement an intrinsic motivation system might be to build a mechanism which can evaluate the degree of novelty of different situations from the point of view of a learning robot, and then designing an associated reward being maximal when these features are in an intermediate level. Maximizing this reward can then create an active exploratory behavior [13, 19].

B. Active imitation learning

A popular social learning modality is imitation learning, also known as learning from demonstration (LfD). It is widely used in robotics as it offers an intuitive framework for non-expert users to (re)program robots. In the context of LfD, active imitation learning [20] is a topic gaining interest. It indeed proposes a possible solution to the problem of choosing what constitutes a good demonstration in terms of generalization capability. The demonstration acquisition being often costly, choosing the demonstrations actively is crucial since it permits to reduce the total number of demonstrations required, and hence the user burden. Active imitation learning has been successfully applied to a variety of robotic tasks such as autonomous navigation [21, 22]. In [23], the authors leverage the uncertainties on a discrete hypothesis space to request meaningful demonstrations from a human teacher. Several approaches have also been proposed in the context where the learner does not request full demonstrations, but only the action to take at a given state [20, 24, 25].

The main component of an active learning framework is a metric allowing to select the demonstration that is expected to yield the greatest improvement. Traditionally, this metric is based on uncertainties [26]. When building statistical models, two different kinds of uncertainties arise: aleatoric and epistemic uncertainties. Aleatoric uncertainties represent the variations in the demonstrations, i.e., different possible ways to achieve a task. This is for example the uncertainty that is captured by probabilistic movement primitives (ProMPs) when fitting a Gaussian or a Gaussian mixture model (GMM) to the demonstrations [27]. Aleatoric uncertainties can typically be employed within a minimal intervention control strategy, where perturbations are corrected only if they have an impact on the task, which results in adaptive tracking gains that take into account the variations of the task [28, 29]. In contrast, epistemic uncertainties represent the uncertainty due to a lack of data, which is crucial information for active learning [26]. In [30], Gaussian process regression (GPR) is used to learn a trajectory given a context. It is applied to a reaching task where the context is the desired end-effector position. Although Gaussian processes are efficient for capturing epistemic uncertainties (model uncertainties), they do not capture aleatoric uncertainties (variations of the task). In [27], an active learning method is proposed for learning movement primitives based on Gaussian mixture models. The context to query (final end-effector position) is selected based on the distance between this context and the different Gaussians of the mixture. In this work, epistemic uncertainties are not considered, and the active learning criterion is based on the aleatoric uncertainties. In contrast, we aim at capturing both types of uncertainties, with an active imitation learning criterion based on epistemic uncertainties. Indeed, work in active learning [26, 31] have shown that metrics based on aleatoric uncertainties are less effective than those based on epistemic uncertainties.

C. Combining intrinsically-motivated learning with social learning

Psychologists have observed on a tool use task that intrinsically-motivated learning can be more efficient if children can see an agent solve the task [12]. This suggests that a learning robot could benefit from combining intrinsically-motivated learning and social learning (e.g., imitation), instead of acquiring skills with a single learning modality. Several works in developmental robotics have indeed studied methods combining those modalities. In [32], Oudeyer *et al.* propose an algorithm for combining intrinsically-motivated self-exploration and imitation learning. In particular, a solution is proposed to the problem of choosing what learning strategy is the most appropriate. In the context of a throwing task, they show that there is a significant gain in combining several learning strategies and actively choosing between them. Besides the fact that their method was only evaluated on a one degree of freedom robot, there is a fundamental difference between their approach and ours. They base the choice of their learning strategy on values of interest levels, which are computed with the progress previously observed when choosing the different modalities. This supposes a notion of competence (reward) to choose between the modalities. In contrast, our

work bases its strategy selection process on uncertainties that are computed with a statistical model representing the data (intrinsically-motivated trials and demonstrations), and hence does not require the notion of an external reward. Additionally, the computation of the interest values in [32] requires the evaluation of the competence before and after each episode, which implies executing a large number of movements to measure the mean distance to the goal. While this does not present any problem in simulation, this is a major drawback for a real robotic application requiring many robot trials. Our method is based on an internal reward related to intrinsic motivation and alleviates therefore this limitation.

An extension of [32] relied on the use of procedures to combine predefined primitive policies [33], but suffers from the same limitations by relying on the same interest model, which does not scale up well to physical robots in the real world. An interest model for goal babbling is also used in [34], by relying on an external reward. In this work, Nguyen *et al.* show that social learning through human demonstrations can bootstrap the performance of an intrinsically motivated robot learner. In a simulated fishing task experiment in which the robot needs to learn how to reach various goals with a fishing rod, a demonstration is given at constant frequency, chosen randomly from the set of goals. They show that this permits to reduce the task cost compared to a purely intrinsically-motivated learning framework. As mentioned in the conclusions of the above papers, an interesting improvement would be to have the possibility to interactively choose the switching between those modalities. Our paper proposes a possible solution to this problem.

III. Bayesian Movement Representation

In this section, we present the movement representation. We build upon the widely used framework of probabilistic movement primitives (ProMPs) [35], which we extend with a Bayesian perspective.

A. Probabilistic movement primitive (ProMP)

A ProMP is a probability distribution over trajectories built from a series of N demonstrations (trajectories) of length T and of D dimensions. A demonstration $\boldsymbol{\tau}_i \in \mathbb{R}^{(T\times D)}$ is approximated by a sum of M basis functions, which are often chosen as radial basis functions (RBF)

$$\boldsymbol{\tau}_i = \boldsymbol{\Phi}\boldsymbol{w}_i + \boldsymbol{\epsilon}, \qquad \text{with} \qquad \boldsymbol{\Phi} = \boldsymbol{\Phi}^{\text{1d}} \otimes \mathbb{I}_D, \tag{1}$$

where $\otimes$ represents the Kronecker product, $\boldsymbol{\epsilon}$ is zero-mean i.i.d. Gaussian noise, $\boldsymbol{w}_i$ of size $MD\times 1$ is the weight associated to the i^{th} demonstration, $\Phi^{\text{1d}}_{T\times M}$ is the basis function matrix with $\Phi^{1d}_{t,m} = \Phi_m(t)$ corresponding to the m^{th} basis function indexed at time t, and $\mathbb{I}_D$ is the identity matrix.

The weight vectors associated to each demonstration are learned through least squares with

$$\boldsymbol{w}_i = (\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\boldsymbol{\tau}_i. \tag{2}$$

A probability distribution $p(\boldsymbol{w})$ can then be learned from the demonstrations $\{\boldsymbol{w}_i\}_{i=1}^N$, usually with a multivariate Gaussian or a GMM.

This probability distribution $p(\boldsymbol{w})$ can then be used for generalization/adaptation to different environments, typically by conditioning on trajectory keypoints.

B. Bayesian Gaussian Mixture Model (BGMM)

In this section, we present the learning of the joint distribution of weights with a BGMM.

1) Joint distribution: The joint distribution is defined by a mixture of K multivariate normal distributions (MVNs) with means $\boldsymbol{\mu}=\{\boldsymbol{\mu}_k\}_{k=1}^K$, precision matrices $\boldsymbol{\Lambda}=\{\boldsymbol{\Lambda}_k\}_{k=1}^K$ and mixing coefficients $\boldsymbol{\pi}=\{\pi_k\}_{k=1}^K$ as

$$p(\boldsymbol{w}|\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda}) = \sum_{k=1}^K \pi_k \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}). \tag{3}$$

A Normal-Wishart prior is used for means and precision matrices, and a Dirichlet prior is put on the mixing coefficients, with

$$p(\boldsymbol{\mu},\boldsymbol{\Lambda}) = \prod_{k=1}^K \mathcal{N}(\boldsymbol{\mu}_k|(\beta_0\boldsymbol{\Lambda}_k)^{-1})\mathcal{W}(\boldsymbol{\Lambda}_k|\boldsymbol{W}_k,\nu_k), \tag{4}$$

$$p(\boldsymbol{\pi}) = \text{Dir}(\boldsymbol{\pi}|\alpha_0). \tag{5}$$

The means, the precision matrices and the mixing coefficients maximizing the posterior distribution are estimated using closed-form update equations similar to those of the Expectation-Maximization (EM) algorithm for the maximum likelihood solution, see Section 10.2.1 in [36] for further details. Also, they are available as parts of standard machine learning libraries (e.g., *scikit-learn* for Python).

Given N demonstrations $\boldsymbol{W} = \{\boldsymbol{w}_i\}_{i=1}^N$, the predictive density of a new weight $\hat{\boldsymbol{w}}$ is equivalent to a mixture of multivariate t-distributions with mean $\hat{\boldsymbol{m}}_k$, covariance matrix $\hat{\boldsymbol{L}}_k$, mixing coefficient $\hat{\pi}_k$ and degrees of freedom $\hat{\nu}_k$ [36], namely

$$p(\hat{\boldsymbol{w}}|\boldsymbol{W}) = \sum_{k=1}^K \hat{\pi}_k \, \text{t}(\hat{\boldsymbol{w}}|\hat{\boldsymbol{m}}_k, \hat{\boldsymbol{L}}_k, \hat{\nu}_k), \tag{6}$$

$$\text{where} \quad \begin{aligned} \hat{\pi}_k &= \frac{\alpha_k}{\sum_{k=1}^K \alpha_k}, \\ \hat{\nu}_k &= \nu_k + 1 - MD, \\ \hat{\boldsymbol{L}}_k &= \frac{(\nu_k+1-MD)\beta_k}{1+\beta_k}\boldsymbol{W}_k, \\ \hat{\boldsymbol{m}}_k &= \boldsymbol{m}_k, \end{aligned} \tag{7}$$

where α_k, β_k and $\boldsymbol{m}_k$ are derived from statistics of the data. We do not include the full equations here, but the reader can refer to Equations (10.41)–(10.63) of [36] for more details.

2) Conditional distribution: The weights represent the evolution of the state with time. For instance, the state can represent the joint angle values of a robot manipulator, or the Cartesian position of an object. We can then condition on a particular value $\hat{\boldsymbol{w}}^i$ of an input dimension (e.g., dimensions representing the object) to get the conditional posterior predictive distribution $p(\hat{\boldsymbol{w}}^o|\hat{\boldsymbol{w}}^i, \boldsymbol{W})$ of an output dimension

(e.g., dimensions representing the robot joint space), as in [36] (Section 10.2.3), namely

$$p(\hat{\boldsymbol{w}}^o|\hat{\boldsymbol{w}}^i,\boldsymbol{W}) = \sum_{k=1}^{K} \hat{\pi}_k^{oi}\ \mathrm{t}(\hat{\boldsymbol{w}}^i|\hat{\boldsymbol{m}}_k^{oi},\hat{\boldsymbol{L}}_k^{oi},\hat{\nu}_k^{oi}), \quad \text{with} \tag{8}$$

$$\hat{\pi}_k^{oi} = \frac{\hat{\pi}_k\ \mathrm{t}(\hat{\boldsymbol{w}}^i|\hat{\boldsymbol{m}}_k^{i},\hat{\boldsymbol{L}}_k^{i},\nu_k^{i})}{\sum_{j=1}^{K}\hat{\pi}_j\ \mathrm{t}(\hat{\boldsymbol{w}}^i|\hat{\boldsymbol{m}}_j^{i},\hat{\boldsymbol{L}}_j^{i},\nu_j^{i})}, \tag{9}$$

$$\hat{\nu}_k^{oi} = \hat{\nu}_k + D^i, \tag{10}$$

$$\hat{\boldsymbol{m}}_k^{oi} = \hat{\boldsymbol{m}}_k^{o} + \hat{\boldsymbol{L}}_k^{oi}\hat{\boldsymbol{L}}_k^{ii^{-1}}(\hat{\boldsymbol{w}}^i - \hat{\boldsymbol{m}}_k^i), \tag{11}$$

$$\hat{\boldsymbol{L}}_k^{oi} = \frac{\hat{\nu}_k + (\hat{\boldsymbol{w}}^i - \hat{\boldsymbol{m}}_k^i)^\top \hat{\boldsymbol{L}}_k^{ii^{-1}}(\hat{\boldsymbol{w}}^i - \hat{\boldsymbol{m}}_k^i)}{\hat{\nu}_k^{oi}} \cdot (\hat{\boldsymbol{L}}_k^{oo} - \hat{\boldsymbol{L}}_k^{oi}\hat{\boldsymbol{L}}_k^{ii^{-1}}\hat{\boldsymbol{L}}_k^{oi^\top}), \tag{12}$$

where we have decomposed $\hat{\boldsymbol{L}}_k = \begin{bmatrix} \hat{\boldsymbol{L}}_k^{ii} & \hat{\boldsymbol{L}}_k^{oi^\top} \\ \hat{\boldsymbol{L}}_k^{oi} & \hat{\boldsymbol{L}}_k^{oo} \end{bmatrix}$.

Due to the linear relation from trajectory space to weight space, a useful property of ProMPs and other trajectory distributions is the possibility to condition on a trajectory via-point/s $\hat{\boldsymbol{\tau}}^i$ directly to get $p(\hat{\boldsymbol{w}}^o|\hat{\boldsymbol{\tau}}^i,\boldsymbol{W})$. This is done simply by replacing all $\hat{\boldsymbol{m}}_k^*$ and $\hat{\boldsymbol{L}}_k^*$ in (8)–(12) by $\boldsymbol{\Phi}\hat{\boldsymbol{m}}_k^*$ and $\boldsymbol{\Phi}\hat{\boldsymbol{L}}_k^*\boldsymbol{\Phi}^\top$, and $\hat{\boldsymbol{w}}^i$ by $\hat{\boldsymbol{\tau}}^i$, respectively.

C. Quantifying the uncertainties

We propose here a method to derive and quantify epistemic uncertainties, which will be the core of our active learning approach. First, we show how aleatoric and epistemic uncertainties can be separated when conditioning. Then, we propose a closed-form information gain metric based on the entropy of the conditional distribution.

1) Uncertainty decomposition: The conditional posterior predictive distribution of the Bayesian ProMP encodes two types of uncertainties: the aleatoric uncertainty (possible variations of the task, the one learned with standard ProMPs) and the epistemic uncertainty (representing the lack of knowledge).

The covariance matrix of the conditional posterior predictive distribution of (12) can be decomposed into aleatoric and epistemic parts [25, 36] as

$$\hat{\boldsymbol{L}}_k^{oi} = \hat{\boldsymbol{L}}_k^{\mathrm{al}} + \hat{\boldsymbol{L}}_k^{\mathrm{ep}}, \quad \text{where} \tag{13}$$

$$\hat{\boldsymbol{L}}_k^{\mathrm{al}} = \frac{\hat{\nu}_k}{\hat{\nu}_k^{oi}}(\hat{\boldsymbol{L}}_k^{oo} - \hat{\boldsymbol{L}}_k^{oi}\hat{\boldsymbol{L}}_k^{ii^{-1}}\hat{\boldsymbol{L}}_k^{oi^\top}), \tag{14}$$

$$\hat{\boldsymbol{L}}_k^{\mathrm{ep}} = \frac{(\hat{\boldsymbol{w}}^i - \hat{\boldsymbol{m}}_k^i)^\top \hat{\boldsymbol{L}}_k^{ii^{-1}}(\hat{\boldsymbol{w}}^i - \hat{\boldsymbol{m}}_k^i)}{\hat{\nu}_k^{oi}}(\hat{\boldsymbol{L}}_k^{oo} - \hat{\boldsymbol{L}}_k^{oi}\hat{\boldsymbol{L}}_k^{ii^{-1}}\hat{\boldsymbol{L}}_k^{oi^\top}). \tag{15}$$

Notice that the aleatoric uncertainty does not depend on the context $\hat{\boldsymbol{w}}^i$, while the epistemic uncertainty grows quadratically with it, which was for example observed in [36] for Bayesian linear regression. Such a decomposition is particularly useful in the context of ProMPs, because we can have access to the aleatoric uncertainty to design minimal intervention control behaviors, or the epistemic uncertainty for quantifying the lack of knowledge of the model.

2) Uncertainty measurement: The most general and common uncertainty measure is the Shannon entropy [36, 37]. Initially proposed for discrete random variables, the Shannon entropy has been extended to continuous probability distributions, in which case it is called continuous (or differential) entropy. We propose to quantify the uncertainty of our conditional ProMP by calculating the (continuous) entropy of its epistemic part.

The entropy of a mixture of multivariate t-distributions cannot be obtained analytically. To avoid computationally expensive Monte Carlo sampling methods, we propose to approximate the distribution with a GMM, for which there is a closed-form lower bound of the entropy. The epistemic part of the conditional ProMP distribution can be approximated by a mixture of K Gaussians using moment-matching:

$$\tilde{\pi}_k(\boldsymbol{c}) = \hat{\pi}_k^{oi}, \qquad \tilde{\boldsymbol{\mu}}_k(\boldsymbol{c}) = \hat{\boldsymbol{m}}_k^{oi}, \qquad \tilde{\boldsymbol{\Sigma}}_k(\boldsymbol{c}) = \frac{\hat{\nu}_k^{oi}}{\hat{\nu}_k^{oi}-2}\hat{\boldsymbol{L}}_k^{\mathrm{ep}}(\boldsymbol{c}). \tag{16}$$

We propose to use the closed-form lower bound introduced in [38], which has been shown to be tight. It is expressed as (to simplify the notation, we omit the fact that all GMM parameters depend on $\boldsymbol{c}$)

$$H_{lower}(p^{ep}(\hat{\boldsymbol{w}}^o|\hat{\boldsymbol{w}}^i,\boldsymbol{W})) = \frac{1}{2}\Big(K\log 2\pi + K + \sum_{i=1}^{K}\tilde{\pi}_i\log|\tilde{\boldsymbol{\Sigma}}_i|\Big) - \sum_{i=1}^{K}\tilde{\pi}_i\log\sum_{j=1}^{K}\tilde{\pi}_j e^{-C_\alpha(p_i,p_j)}, \tag{17}$$

where $C_\alpha(p_i,p_j)$ is the Chernoff α-divergence distance function between i^{th} and j^{th} Gaussians for $\alpha\in[0,1]$:

$$C_\alpha(p_i,p_j) = \frac{(1-\alpha)\alpha}{2}(\tilde{\boldsymbol{\mu}}_i - \tilde{\boldsymbol{\mu}}_j)^\top\Big((1-\alpha)\tilde{\boldsymbol{\Sigma}}_i + \alpha\tilde{\boldsymbol{\Sigma}}_j\Big)^{-1}(\tilde{\boldsymbol{\mu}}_i - \tilde{\boldsymbol{\mu}}_j) + \frac{1}{2}\log\left(\frac{|(1-\alpha)\tilde{\boldsymbol{\Sigma}}_i + \alpha\tilde{\boldsymbol{\Sigma}}_j|}{|\tilde{\boldsymbol{\Sigma}}_i|^{1-\alpha}|\tilde{\boldsymbol{\Sigma}}_j|^{\alpha}}\right). \tag{18}$$

In practice we choose $\alpha = 1/2$, in which case the Chernoff divergence corresponds to the Bhattacharyya distance[1].

We will now show how we can use the learned statistical model to build different active learning modalities.

IV. Active learning modalities

In this section, we derive two active learning strategies from the learned joint model: imitation and intrinsically-motivated learning, and a criterion for choosing which learning modality is better suited at the current learning stage. To facilitate the presentation of the approach, we will introduce the approach in the context of a specific robot experiment, where the aim is to learn to move an object to different positions. First, we present the task and the goal of the active learning framework. Secondly, we present the proposed method for active imitation

[1]It is shown in [38] that choosing $\alpha = 1/2$ gives the tightest lower bound for a homoscedastic mixture of Gaussians. In our heteroscedastic case, one could optimize over α to find the tightest lower bound, but we observed that this did not yield any improvement in practice, and hence fixed $\alpha = 1/2$.

learning. Then, we propose a method for active intrinsically-motivated learning. Finally, a criterion for actively choosing whether imitation or intrinsically-motivated learning is better suited is presented.

A. Manipulation task

We present our approach in the context of learning to manipulate an object with a robot. The trajectory is composed of the robot joint states $\boldsymbol{\tau}^{\text{robot}}$ and the object position $\boldsymbol{\tau}^{\text{obj}}$, which implies that the ProMP weights $\boldsymbol{w}$ are a concatenation of robot weights $\boldsymbol{w}^{\text{robot}}$ and object weights $\boldsymbol{w}^{\text{obj}}$.

The goal of the task is to move the object to different desired final object positions $\boldsymbol{\tau}_{\text{des}}^{\text{obj},t=T}$. We denote the goal space $\mathcal{G}$ as the space of all desired final object positions we would like our robot to be able to generalize to. Formally, this means that there exists an unknown ground truth target distribution $p^{\text{GT}}(\boldsymbol{w}) = p^{\text{GT}}(\boldsymbol{w}^{\text{robot}}, \boldsymbol{w}^{\text{obj}})$ which can be used to generate robot movements $p^{\text{GT}}(\boldsymbol{w}^{\text{robot}}|\boldsymbol{\tau}_{\text{des}}^{\text{obj},t=T})$ that bring the object to the position $\boldsymbol{\tau}_{\text{des}}^{\text{obj},t=T}$.

We aim to learn this unknown joint distribution by combining imitation and intrinsically-motivated learning.

B. Imitation learning

We suppose here that there exists a human demonstrator/oracle that can be queried to demonstrate a robot movement that brings the object to any desired final position $\boldsymbol{\tau}_{\text{des}}^{\text{obj},t=T}$ in $\mathcal{G}$. Acquiring these demonstrations is usually cumbersome, therefore we would like the demonstrations to be as informative as possible. We propose to choose the demonstration with active learning to alleviate this limitation.

Given a current database of movements $\boldsymbol{W}$, we propose to leverage the uncertainties learned by the BGMM and choose the goal $\boldsymbol{\tau}_{\text{des}}^{\text{obj},t=T}$ for which the entropy of the epistemic part of the conditional distribution $p(\boldsymbol{w}^{\text{robot}}|\boldsymbol{\tau}_{\text{des}}^{\text{obj},t=T}, \boldsymbol{W})$ is maximal. As explained in the previous section, this entropy is not easy to compute for GMMs, so we instead maximize a closed-form lower bound. The full active imitation learning algorithm is shown in Algorithm 1.

C. Intrinsically-motivated learning

We present here another learning modality, where the robot can try out a movement by itself and observe the environment changes in an open-ended manner. Namely, the robot chooses to execute a particular movement and observes the movement of the object. In contrast to imitation learning, one major advantage of intrinsically-motivated learning is that it does not require the presence of a human demonstrator.

We propose to select a robot movement based on how uncertain we are about the object movements it will cause. Formally, we would like to try the robot movement that maximizes the entropy of the epistemic part of the conditional distribution $p(\boldsymbol{w}^{\text{obj}}|\boldsymbol{w}^{\text{robot}}, \boldsymbol{W})$, but this poses several problems. From a robotics point of view, doing so might pose safety problems as the movement retrieved might be very far from the underlying distribution $p^{\text{GT}}(\boldsymbol{w}^{\text{robot}})$ we aim to learn. From an active learning point of view, our active learning selection scheme is myopic and such criterion might select robot movements far away from the underlying distribution, i.e., where no generalization is required. For these reasons, we propose to use an information-density method [36]. Namely, we aim to find a robot movement that both has high information content (in the sense of the epistemic entropy), and that is close to the distribution of robot movements $p^{\text{robot}}(\boldsymbol{w}^{\text{robot}}|\boldsymbol{W})$:

$$\boldsymbol{w}^{\text{robot}^*} = \arg\max_{\boldsymbol{w}^{\text{robot}} \in \mathcal{W}^{\text{robot}}} \Big[H_{lower}\Big(p^{ep}(\boldsymbol{w}^{\text{obj}}|\boldsymbol{w}^{\text{robot}}, \boldsymbol{W})\Big) + \beta p^{\text{robot}}(\boldsymbol{w}^{\text{robot}}) \Big], \tag{19}$$

where β is an hyperparameter weighting the relative importance of the two costs.

The full intrinsically-motivated learning algorithm is shown in Algorithm 2.

D. Choosing the learning modality

We have presented two different learning modalities: imitation learning and intrinsically-motivated learning[2]. We propose here a method to choose between these learning modalities.

A difficulty in choosing the right learning modality is that the epistemic entropies are not comparable for the two learning modalities. Indeed, for imitation learning we focus on the epistemic entropy of the robot movement conditional distribution for a given object final position, whereas for intrinsically-motivated learning we look at the epistemic entropy of the object movement conditional distribution for a given robot movement.

We propose to compare these learning modalities in terms of the expected reduction of the epistemic entropies of the robot movement given the desired goal. This means that we aim to

Algorithm 1: Active imitation learning

Data: Movement database $\boldsymbol{W} = \{\boldsymbol{w}_i^{\text{robot}}, \boldsymbol{w}_i^{\text{obj}}\}_{i=1}^N$, goal space $\mathcal{G}$
Result: goal $\boldsymbol{\tau}_{\text{des}^*}^{\text{obj},t=T}$ at which to request a demonstration

Learn joint distribution of $p(\boldsymbol{w}|\boldsymbol{W}) = p(\boldsymbol{w}^{\text{robot}}, \boldsymbol{w}^{\text{obj}}|\boldsymbol{W})$ with BGMM;
Calculate $p(\boldsymbol{w}^{\text{robot}}|\boldsymbol{\tau}_{\text{des}}^{\text{obj},t=T}, \boldsymbol{W})$ using Eqs (8) to (12);
Isolate the epistemic uncertainty $p^{ep}(\boldsymbol{w}^{\text{robot}}|\boldsymbol{\tau}_{\text{des}}^{\text{obj},t=T}, \boldsymbol{W})$ with Eqs (13) and (15);
Approximate the entropy of $p^{ep}(\boldsymbol{w}^{\text{robot}}|\boldsymbol{\tau}_{\text{des}}^{\text{obj},t=T}, \boldsymbol{W})$ with Eqs (16) to (18);

Find $\boldsymbol{\tau}_{\text{des}^*}^{\text{obj},t=T} = \arg\max_{\boldsymbol{\tau}_{\text{des}}^{\text{obj}}, \boldsymbol{t=T} \in \mathcal{G}} [H_{lower}(p^{ep}(\boldsymbol{w}^{\text{robot}}|\boldsymbol{\tau}_{\text{des}}^{\text{obj},t=T}, \boldsymbol{W}))]$.

[2]Note that both modalities are based on the same joint model of the movements that has been learned using a BGMM. What changes in those scenarios is the input on which we condition, which can be the desired final object position or the robot movement.

Algorithm 2: Active intrinsically-motivated learning

Data: Movement database $\boldsymbol{W} = \{\boldsymbol{w}_i^{\text{robot}}, \boldsymbol{w}_i^{\text{obj}}\}_{i=1}^N$, robot movement space $\mathcal{W}^{\text{robot}}$
Result: robot movement $\boldsymbol{w}^{\text{robot}^*}$ to execute

Learn joint distribution of $p(\boldsymbol{w}|\boldsymbol{W}) = p(\boldsymbol{w}^{\text{robot}}, \boldsymbol{w}^{\text{obj}}|\boldsymbol{W})$ with BGMM;
Calculate $p(\boldsymbol{w}^{\text{obj}}|\boldsymbol{w}^{\text{robot}}, \boldsymbol{W})$ using Eqs (8) to (12);
Isolate the epistemic uncertainty $p^{ep}(\boldsymbol{w}^{\text{obj}}|\boldsymbol{w}^{\text{robot}}, \boldsymbol{W})$ with Eqs (13) and (15);
Approximate the entropy of $p^{ep}(\boldsymbol{w}^{\text{obj}}|\boldsymbol{w}^{\text{robot}}, \boldsymbol{W})$ with Eqs (16) to (18);
Get the marginal distribution $p^{\text{robot}}(\boldsymbol{w}^{\text{robot}}|\boldsymbol{W})$ from $p(\boldsymbol{w}|\boldsymbol{W})$;

Find $\boldsymbol{w}^{\text{robot}^*} = \arg\max_{\boldsymbol{w}^{\text{robot}} \in \mathcal{W}^{\text{robot}}} [H_{lower}(p^{ep}(\boldsymbol{w}^{\text{obj}}|\boldsymbol{w}^{\text{robot}}, \boldsymbol{W})) + \beta p^{\text{robot}}(\boldsymbol{w}^{\text{robot}})]$.

minimize the expected (over the goal space) epistemic entropy on the robot movement when conditioning on the desired goal. This notion of expected epistemic entropy corresponds to

$$EE\Big(\boldsymbol{W}\Big) = \mathbb{E}_{\boldsymbol{\tau}_{\text{des}}^{\text{obj}}, \boldsymbol{t=T} \in \mathcal{G}}\Big[H_{lower}\Big(p^{ep}(\boldsymbol{w}^{\text{robot}}|\boldsymbol{\tau}_{\text{des}}^{\text{obj},t=T}, \boldsymbol{W})\Big)\Big]. \tag{20}$$

This expected epistemic entropy permits us to introduce the notion of expected epistemic entropy reduction, which is the reduction of the expected epistemic entropy when adding a datapoint $\boldsymbol{w}_{\text{new}}$ to the database $\boldsymbol{W}$:

$$EER\Big(\boldsymbol{w}_{\text{new}}\Big|\boldsymbol{W})\Big) = EE\Big(\boldsymbol{W}\Big) - EE\Big(\boldsymbol{W} \cup \{\boldsymbol{w}_{\text{new}}\}\Big). \tag{21}$$

In practice, computing the expected epistemic entropy reduction involves the relearning of the BGMM with the augmented dataset $\boldsymbol{W} \cup \{\boldsymbol{w}_{\text{new}}\}$ and computing the expected epistemic entropy on this new joint model. This notion of expected epistemic entropy can straightforwardly be extended to a distribution[3] of potential new datapoints $p_{\text{new}}(\boldsymbol{w})$ with

$$EER\Big(p_{\text{new}}(\boldsymbol{w})\Big|\boldsymbol{W}\Big) = \mathbb{E}_{\boldsymbol{w}_{\text{new}} \sim p_{\text{new}}} EER\Big(\boldsymbol{w}_{\text{new}}\Big|\boldsymbol{W}\Big). \tag{22}$$

We will show now how we can use this to calculate the expected reduction of epistemic entropy when choosing imitation learning or intrinsically-motivated learning.

a) Imitation learning: Algorithm 1 returns the goal $\boldsymbol{\tau}_{\text{des}^*}^{\text{obj},t=T}$ that should yield the most informative demonstration. Even though we do not know in advance what demonstration $\boldsymbol{w}_{\text{new}}$ we will get when querying the demonstrator, we can use our model to compute the distribution of potential demonstrations $p(\boldsymbol{w}_{\text{new}}|\boldsymbol{\tau}_{\text{des}^*}^{\text{obj},t=T}, \boldsymbol{W})$ bringing the object to the desired goal. This allows us to compute the expected epistemic entropy reduction if choosing the imitation learning strategy with

$$EER(\text{Imitation}) = EER\Big(p(\boldsymbol{w}_{\text{new}}|\boldsymbol{\tau}_{\text{des}^*}^{\text{obj},t=T}, \boldsymbol{W})\Big|\boldsymbol{W}\Big). \tag{23}$$

b) Intrinsically-motivated learning: Similarly, Algorithm 2 returns the robot movement $\boldsymbol{w}^{\text{robot}^*}$ expected to show an interesting object movement. We can also estimate the expected trajectories $p(\boldsymbol{w}_{\text{new}}|\boldsymbol{w}^{\text{robot}^*}, \boldsymbol{W})$ when executing this robot movement. From this distribution, we compute the expected epistemic entropy reduction if choosing intrinsically-motivated learning with

$$EER(\text{Intrinsic}) = EER\Big(p(\boldsymbol{w}_{\text{new}}|\boldsymbol{w}^{\text{robot}^*}, \boldsymbol{W})\Big|\boldsymbol{W}\Big). \tag{24}$$

In the above, we have proposed a measure to quantify the informativeness of the different learning strategies, which we can use to choose the most appropriate strategy by selecting the one which leads the highest expected epistemic entropy reduction. The selection process of the best learning strategy is summarized in Algorithm 3.

Algorithm 3: Choice of learning strategy

Data: Movement database $\boldsymbol{W} = \{\boldsymbol{w}_i^{\text{robot}}, \boldsymbol{w}_i^{\text{obj}}\}_{i=1}^N$, goal space $\mathcal{G}$, robot movement space $\mathcal{W}^{\text{robot}}$
Result: the learning strategy (*Imitation* or *Intrinsically-motivated*) that is better suited

Find $\boldsymbol{\tau}_{\text{des}^*}^{\text{obj},t=T}$ with Alg.1;
Compute the expected epistemic uncertainty reduction of imitation learning $EER(\text{Imitation})$ with Eq.23;

Find $\boldsymbol{w}^{\text{robot}^*}$ with Alg.2;
Compute the expected epistemic uncertainty reduction of intrinsically-motivated learning $EER(\text{Intrinsic})$ with Eq.24;

if $EER(\textit{Imitation}) > EER(\textit{Intrinsic})$ **then**
  Return *Imitation*
else
  Return *Intrinsically-motivated*
end

V. EXPERIMENTS

In this section, we show the usefulness of our approaches in the context of a robotic task. First, we present the waste throwing task we consider. Then, we evaluate quantitatively the performance of our approaches for imitation learning, intrinsically-motivated learning, and the combination of both.

A. Waste throwing task

We consider the task of throwing waste with a 7 DoF Franka Emika Panda robot simulated in pyBullet [39]. This task is essential for the broader challenge of automatizing various forms of recycling. It is also relevant in diverse industrial

[3]In practice for computational reasons, we approximate $EER\Big(p_{\text{new}}(\boldsymbol{w})\Big|\boldsymbol{W}\Big)$ by $EER\Big(\boldsymbol{w}_{\text{new}}^{\text{MP}}\Big|\boldsymbol{W})\Big)$, where $\boldsymbol{w}_{\text{new}}^{\text{MP}}$ denotes the most probable datapoint under $p_{\text{new}}(\boldsymbol{w})$.

applications requiring a robot to sort objects fast within a limited workspace.

An overview of the simulated setup can be seen in Fig. 1. The goal of the task is to be able to generate robot movements that bring a simulated can to different desired positions within a goal space $\mathcal{G}$. The particularity of this goal space is that, for a part of it, it is possible to bring the object with a non-dynamic movement because the desired final position is in the reachable robot workspace. However, for the rest of the goal space, the final desired object position is outside of the robot workspace, so that it requires the robot to throw the can with a dynamic movement. For benchmarking and reproducibility purposes, we build our experiments on a precomputed database of demonstrations. We create 200 non-dynamic demonstrations and 260 dynamic demonstrations using an oracle, that we gather in a database of demonstrations $\mathcal{D}$. In Fig. 1, we illustrate the can trajectory for three dynamic demonstrations and three non-dynamic demonstrations. In Fig. 2, we show the final can positions in our database, with the blue color representing the non-dynamic demonstrations and the orange color representing the dynamic demonstrations.

The trajectories of our database encode the robot movement at a frequency of 240Hz, with $T = 639$ timesteps, representing movements of about 3 seconds. We choose a 10-dimensional state space containing the 7 joint angle values of the robot, and the 3-dimensional Cartesian position of the can. In all experiments, we use $N = 30$ Gaussian radial basis functions[4] (RBFs) for ProMP. The width of the RBFs are set as $h = (\frac{T-1}{N})^2$, and the centers $\{c_m\}_{m=1}^{D}$ are evenly spaced between $-2h$ and $T+2h$. We choose a diagonal covariance matrix prior, with a standard deviation of 0.1 for the ProMP weights, and a mean concentration prior of 0.0001. We use a maximum number of 5 Gaussians, or strictly less than the number of demonstrations if there are less than 6 demonstrations. Other hyperparameters of the BGMM are the default hyperparameters of the *scikit-learn* library [40].

The maximization procedure in active imitation learning and active intrinsically-motivated learning is performed using a Bayesian optimization algorithm: the Tree-Structured Parzen Estimator approach (TPE) [41], implemented in the Python package hyperopt [42]. A maximal number of iterations of 100 is used in the algorithm. For imitation learning, we use a 2-dimensional uniform search space corresponding to the goal space. For intrinsically-motivated learning, as the space of possible robot movements is of high dimension (30 basis functions $\times$ 7 joint angles), we perform the search on the first two principal components of $\{\boldsymbol{w}_i^{\text{robot}}\}_{i=1}^{N}$, found by principal component analysis (PCA) [43]. The search space that we use is then the marginal distribution $p(\boldsymbol{w}^{\text{robot}})$ projected to the 2-dimensional PCA subspace.

We introduce an objective metric for comparing our learning modalities: the task cost, which is simply a ℓ_2 norm between the final object position and the desired object position, averaged over the goal space. In practice, we compute this task

[4]Namely: $\Phi_m(t) = \frac{\phi_m(t)}{\sum_{n=1}^{D}\phi_n(t)}$ with $\phi_m(t) = \exp\left(-\frac{(t-c_m)^2}{2h}\right)$.

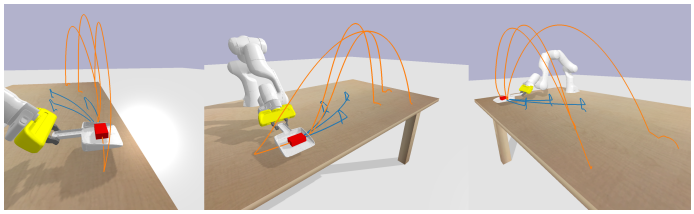

Fig. 1: Object trajectory for 6 demonstrations of the database (3 dynamic demonstrations in orange, and 3 non-dynamic demonstrations in blue).

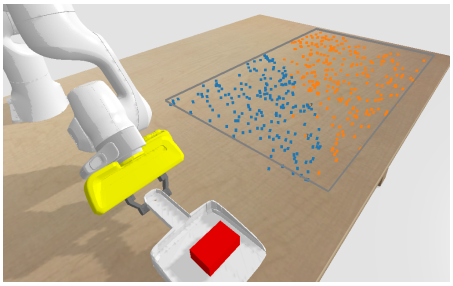

Fig. 2: Desired final object positions. The grey rectangle represents the goal space $\mathcal{G}$. Blue/orange dots show the final object position of respectively the non-dynamic/dynamic demonstrations of the database.

cost by computing the maximum *a posteriori* robot movement given a goal chosen over a uniform grid of 5×5 goals in the goal space, execute those 25 movements in simulation, and average the ℓ_2 norms between the final object positions and the desired object positions. Such a metric presents the advantage of being directly representative of the quality of the learned task, while remaining agnostic to the metrics we chose for active learning. It is important to note here that this metric based on an external reward is used only for comparison, and not by our active learning algorithms.

B. Imitation learning

We present here the results of our method in an imitation learning scenario.

First, we show qualitatively in Fig. 3 our method during 20 iterations of active learning, starting with 2 random initial demonstrations. We can see in this figure that our method effectively selects goals that are far from goals already observed in available demonstrations. Now, we propose to evaluate our method quantitatively. We benchmark our method against two different active learning baselines:

- Random: this baseline simply selects a random goal g from $\mathcal{G}$.
- Minimum likelihood (Min. Lik.): this method, similar to [27], chooses the goal that is the furthest from our current task representation. Formally, this means that we compute the marginal distribution of our BGMM over the goal space, and choose the goal that has the minimum likelihood under this distribution.

We initialize the learning process with 2 initial demonstrations randomly sampled from the database. For our method and

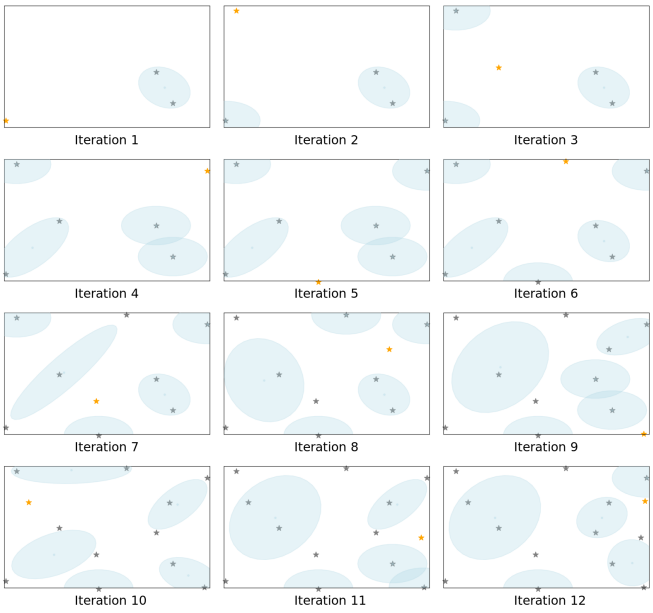

Fig. 3: Evolution of the active imitation learning strategy. The goal space is represented in this figure. Grey stars represent the final object position of the available demonstrations, and orange stars the selected goal to query. The transparent ellipses show the marginal distribution of the BGMM on the goal space.

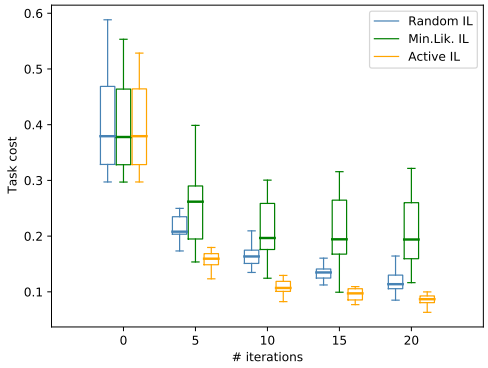

Fig. 4: Evaluation of imitation learning strategy.

the baselines, the experiment is reproduced 20 times, starting from different initial demonstrations. The results are shown in Fig. 4. We can see that our method outperforms both baselines in terms of task cost reduction across the learning process. Notably, it performs around 30% better than the random strategy at all stages of the learning process (at 5, 10, 15, and 20 iterations), and about 50% better than the minimum likelihood strategy. This shows that the epistemic uncertainty seems to be a good criterion for goal selection. Also, it confirms the usefulness of this low-level arbitration capability deciding where the agent currently needs to request a demonstration.

C. Intrinsically-motivated learning

We present here the results of our intrinsically-motivated learning method. First, we would like to emphasize quantitatively the need for combining imitation learning and intrinsically-motivated learning for this waste throwing task. Namely, we want to show that using intrinsically-motivated learning can effectively reduce the task cost. We show in Fig. 5 the task cost (averaged over 20 demonstrations) for:

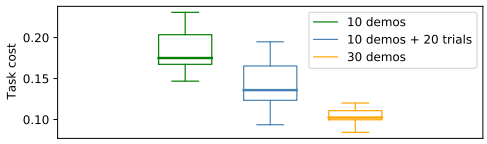

Fig. 5: Influence of demonstrations for intrinsically-motivated learning strategy.

- 10 random demonstrations;
- 10 random demonstrations + 20 active intrinsically-motivated trials;
- 30 random demonstrations.

We can see that, starting from 10 initial demonstrations, 20 intrinsically-motivated learning trials can improve the model. We can notably see that 20 intrinsically-motivated trials reduce the task cost half as well as 20 additional demonstrations. This shows that intrinsically-motivated learning can be used to reduce the burden of the human demonstrator by reducing the number of demonstrations s/he will be asked. Namely, Fig. 5 shows that intrinsically-motivated learning seems to be a good learning modality to be combined with imitation learning. We propose now a baseline to compare our intrinsically-motivated learning method with:

- Random: This baseline computes the marginal $p(\boldsymbol{w}_{\text{robot}}|\boldsymbol{W})$ from the BGMM, and samples a robot movement from it. This seems like a reasonable baseline which already uses the correlations in the observed robot movements, and samples meaningful robot movements that are close to the observed demonstrations.

In Fig. 6, we show the performance of our method compared to this baseline, averaged over 20 experiments, and starting from 5 or 10 randomly sampled initial demonstrations. We can observe that our method presents a clear improvement over the baseline in both cases. Namely, the baseline deteriorates the task cost across the iterations, whereas our method permits to reduce the task cost, as observed in Fig. 5 (the mean task cost is reduced by around 20% after 10 autonomous trials in both cases). The deterioration of the task cost with the random approach can be explained by the fact that sampling from the marginal distribution of the robot movements at each iteration might end up with samples that are quite far from the original distribution, hence not useful for the task.

D. Choice of learning modality

Here, we show the usefulness of choosing actively the learning modality at each iteration of the learning process. Our results, averaged over 20 experiments, start with 2 initial demonstrations (randomly sampled). In Fig. 7, we show which learning modalities are chosen by our method during the learning process. We can see that, for the first 5 iterations, the

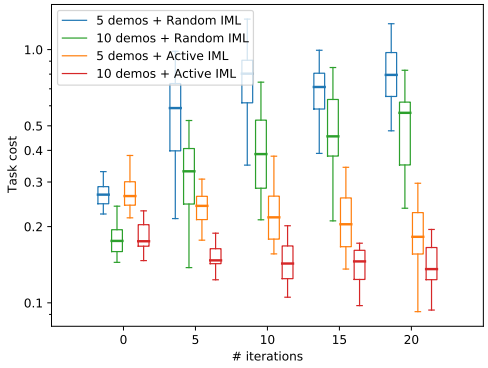

Fig. 6: Evaluation of intrinsically-motivated learning strategy (task cost in logarithmic scale).

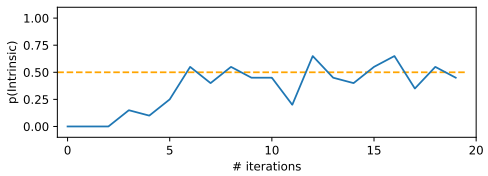

Fig. 7: Example of a learning process in which the learning strategy is selected at each step based on the proposed active learning method.

imitation learning strategy is almost always preferred, while afterwards the two learning modalities are selected with about the same probability. On average, the intrinsically-motivated learning modality is chosen with a probability of 36%. Leveraging this knowledge, we introduce a baseline which simply chooses the intrinsically-motivated learning strategy in a random manner with a probability 0.36, and imitation learning otherwise. Note that this baseline is already quite good, as it involves the information of the optimal probability of selecting the intrinsically-motivated strategy obtained with our method. The results are shown in Fig. 8. We observe that our method outperforms this baseline in the beginning of the learning process (at iteration 5), but gives similar results later in the training process. This suggests that our method for choosing the learning modality is useful for the investigated task, especially in the beginning of the learning process. In Fig. 8, we also show the performance of two additional baselines choosing always the same learning modality. We can see that choosing always *intrinsically-motivated learning* results in very poor learning. This is because two initial demonstrations are not sufficient to be able to generate meaningful movement variations. This is consistent with the fact that *imitation learning* should be preferred in the beginning of the learning process, as our method has automatically discovered (see Fig. 7). We also observe in Fig. 8 that choosing actively the learning modalities results in a task cost on par with only *imitation learning* across the whole learning process, which is a nice result because it means that we can reduce the number of demonstrations by 36% without suffering from a performance degradation, and therefore reduce the human

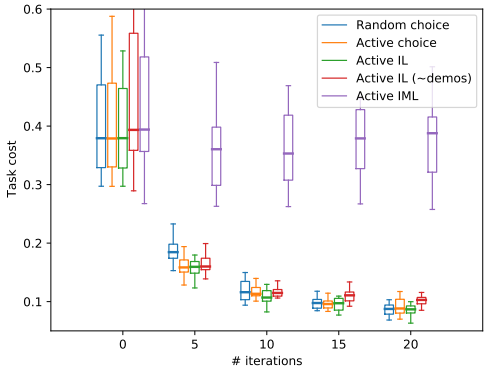

Fig. 8: Evaluation for the choice of the learning strategy.

burden of providing demonstrations. A fairer comparison is to compare our method against only *imitation learning* with the same number of demonstrations [5], which we also plotted in Fig. 8. We can see that our method outperforms this baseline at iterations 15 and 20 by around 15%. This therefore motivates the meta-level arbitration capability of our framework for orchestrating the different learning modalities.

VI. Conclusion

In this article, we proposed a Bayesian representation of robot movements by extending the widely-used framework of probabilistic movement primitives. With this Bayesian representation, we proposed three active learning criteria leveraging the knowledge of the model uncertainties (epistemic uncertainties) that permit two different learning modalities (imitation learning and intrinsically-motivated learning) as well a principled method for arbitrating between them in an open-ended manner. To the best of our knowledge, our work is the first to integrate those three aspects.

We showed the robustness of our approach with a waste throwing task with a 7-DoF simulated Franka Emika Panda robot. We studied the usefulness of each of our active learning algorithms by comparing them to alternative baselines, and showed that in all experiments, our algorithms give the best performance.

The fundamental element of our method lies in that we model the joint distribution of the movement. By doing so, we can compute several forms of conditional distributions (in our case, quantifying the effect of a specific robot movement on the object for intrinsically-motivated learning, or the robot movement needed to bring the object to a desired final position for imitation learning). Also, as intrinsically-motivated learning and imitation learning are based on the same joint model of the movement, we have shown that we can compare these very different learning modalities quantitatively.

In future work, we will study whether additional learning modalities can be added to the framework. In particular, the use of human feedback as a learning modality could be particularly interesting as it would be less cumbersome for the human user to give the robot partial feedback rather than full

[5]Namely 0, 5, 8, 10, 13 demonstrations at iterations 0, 5, 10, 15, 20.

demonstrations. More generally, we will also investigate if the proposed active learning approach can be extended to different aspects of the skills, in order to allow the different learning modalities to improve different aspects of the task (e.g., acquiring the kinematics aspects through observation learning, and the dynamical aspects through experiential learning).

ACKNOWLEDGMENTS

This work was supported by the Swiss National Science Foundation through the ROSALIS project.

REFERENCES

[1] M. Lungarella, G. Metta, R. Pfeifer, and G. Sandini, "Developmental robotics: a survey," *Connection science*, vol. 15, no. 4, pp. 151–190, 2003.

[2] C. J. Charpentier, K. Iigaya, and J. P. O'Doherty, "A neuro-computational account of arbitration between choice imitation and goal emulation during human observational learning," *Neuron*, 2020.

[3] V. Horner and A. Whiten, "Causal knowledge and imitation/emulation switching in chimpanzees (pan troglodytes) and children (homo sapiens)," *Animal cognition*, vol. 8, no. 3, pp. 164–181, 2005.

[4] A. Whiten, N. McGuigan, S. Marshall-Pescini, and L. M. Hopper, "Emulation, imitation, over-imitation and the scope of culture for child and chimpanzee," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 1528, pp. 2417–2428, 2009.

[5] G. Gergely and G. Csibra, "Sylvia's recipe: The role of imitation and pedagogy in the transmission of human culture," in *Roots of Human Sociality: Culture, Cognition, and Human Interaction*, N. J. Enfield and S. C. Levinson, Eds. Berg Publishers, 2006, pp. 229–255.

[6] P. Zukow-Goldring and M. A. Arbib, "Affordances, effectivities and assisted imitation: Caregivers and the directing of attention," *Neurocomputing*, vol. 70, no. 13-15, pp. 2181–2193, 2007.

[7] K. J. Rohlfing, J. Fritsch, B. Wrede, and T. Jungmann, "How can multimodal cues from child-directed interaction reduce learning complexity in robots?" *Advanced Robotics*, vol. 20, no. 10, pp. 1183–1199, 2006.

[8] J. Saunders, C. L. Nehaniv, K. Dautenhahn, and A. Alissandrakis, "Self-imitation and environmental scaffolding for robot teaching," *Intl Journal of Advanced Robotics Systems*, vol. 4, no. 1, pp. 109–124, 2007.

[9] S. Calinon and A. G. Billard, "What is the teacher's role in robot programming by demonstration? - Toward benchmarks for improved learning," *Interaction Studies*, vol. 8, no. 3, pp. 441–464, 2007.

[10] A. L. Thomaz and C. Breazeal, "Teachable robots: Understanding human teaching behavior to build more effective robot learners," *Artificial Intelligence*, vol. 172, pp. 716–737, April 2008.

[11] M. Cakmak, N. DePalma, R. I. Arriaga, and A. L. Thomaz, "Exploiting social partners in robot learning," *Autonomous Robots*, vol. 29, no. 3-4, pp. 309–329, 2010.

[12] L. M. Hopper, E. G. Flynn, L. A. Wood, and A. Whiten, "Observational learning of tool use in children: Investigating cultural spread through diffusion chains and learning mechanisms through ghost displays," *Journal of experimental child psychology*, vol. 106, no. 1, pp. 82–97, 2010.

[13] P.-Y. Oudeyer, F. Kaplan, and V. V. Hafner, "Intrinsic motivation systems for autonomous mental development," *IEEE transactions on evolutionary computation*, vol. 11, no. 2, pp. 265–286, 2007.

[14] J. Schmidhuber, "Formal theory of creativity, fun, and intrinsic motivation (1990–2010)," *IEEE Transactions on Autonomous Mental Development*, vol. 2, no. 3, pp. 230–247, 2010.

[15] R. W. White, "Motivation reconsidered: The concept of competence," *Psychological review*, vol. 66, no. 5, p. 297, 1959.

[16] E. L. Deci and R. M. Ryan, *Intrinsic Motivation and Self-Determination in Human Behavior*. Springer US, 1985.

[17] D. E. Berlyne, *Conflict, arousal, and curiosity.* McGraw-Hill Book Company, 1960.

[18] J. C. Horvitz, "Mesolimbocortical and nigrostriatal dopamine responses to salient non-reward events," *Neuroscience*, vol. 96, no. 4, pp. 651–656, 2000.

[19] J. Marshall, D. Blank, and L. Meeden, "An emergent framework for self-motivation in developmental robotics," *International Conference on Development and Learning*, 2004.

[20] A. Shon, D. Verma, and R. Rao, "Active imitation learning," in *Proc. AAAI Conference on Artificial Intelligence*, 2007.

[21] D. Silver, J. Bagnell, and A. Stentz, "Active learning from demonstration for robust autonomous navigation," in *Proc. IEEE Intl Conf. on Robotics and Automation (ICRA)*, 2012.

[22] O. Kroemer, R. Detry, J. Piater, and J. Peters, "Combining active learning and reactive control for robot grasping," *Robotics and Autonomous systems*, vol. 58, no. 9, pp. 1105–1116, 2010.

[23] C. Chao, M. Cakmak, and A. Thomaz, "Transparent active learning for robots," in *Proc. ACM/IEEE Intl Conf. on Human-Robot Interaction (HRI)*, 2010.

[24] S. Chernova and M. Veloso, "Interactive policy learning through confidence-based autonomy," *Journal of Artificial Intelligence Research*, vol. 34, pp. 1–25, 2009.

[25] H. Girgin, E. Pignat, N. Jaquier, and S. Calinon, "Active improvement of control policies with Bayesian Gaussian mixture model," in *Proc. IEEE/RSJ Intl Conf. on Intelligent Robots and Systems (IROS)*, 2020.

[26] B. Settles, "Active learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 6, no. 1, pp. 1–114, 2012.

[27] A. Conkey and T. Hermans, "Active learning of probabilistic movement primitives," in *Proc. IEEE Intl Conf. on Humanoid Robots (Humanoids)*, 2019, pp. 1–8.

[28] E. Todorov and M. I. Jordan, "A minimal intervention principle for coordinated movement," in *Advances in Neural Information Processing Systems (NIPS)*, 2002, pp. 27–34.

[29] S. Calinon, D. Bruno, and D. G. Caldwell, "A task-parameterized probabilistic model with minimal intervention control," in *Proc. IEEE Intl Conf. on Robotics and Automation (ICRA)*, Hong Kong, China, May-June 2014, pp. 3339–3344.

[30] G. Maeda, G. Neumann, M. Ewerton, R. Lioutikov, O. Kroemer, and J. Peters, "Probabilistic movement primitives for coordination of multiple human–robot collaborative tasks," *Autonomous Robots*, vol. 41, no. 3, pp. 593–612, 2017.

[31] T. Kulak, H. Girgin, and S. Odobez, J.-M.and Calinon, "Active learning of Bayesian probabilistic movement primitives," *IEEE Robotics and Automation Letters*, 2021.

[32] S. M. Nguyen and P.-Y. Oudeyer, "Active choice of teachers, learning strategies and goals for a socially guided intrinsic motivation learner," *Paladyn*, vol. 3, no. 3, pp. 136–146, 2012.

[33] N. Duminy, S. M. Nguyen, and D. Duhaut, "Learning a set of interrelated tasks by using a succession of motor policies for a socially guided intrinsically motivated learner," *Frontiers in neurorobotics*, vol. 12, p. 87, 2019.

[34] S. M. Nguyen, A. Baranes, and P.-Y. Oudeyer, "Bootstrapping intrinsically motivated learning with human demonstrations," *Proc. IEEE Intl Conf. on Development and Learning (ICDL)*, pp. 1–8, 2011.

[35] A. Paraschos, C. Daniel, J. R. Peters, and G. Neumann, "Probabilistic movement primitives," in *Advances in Neural Information Processing Systems (NIPS)*, 2013, pp. 2616–2624.

[36] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, 2006.

[37] C. Shannon, "A mathematical theory of communication," *Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[38] A. Kolchinsky and B. Tracey, "Estimating mixture entropy with pairwise distances," *Entropy*, vol. 19, no. 7, p. 361, 2017.

[39] E. Coumans and Y. Bai, "Pybullet, a python module for physics simulation for games, robotics and machine learning," http://pybullet.org, 2016–2020.

[40] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[41] J. S. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyper-parameter optimization," in *Advances in neural information processing systems*, 2011, pp. 2546–2554.

[42] J. Bergstra, D. Yamins, and D. D. Cox, "Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms," in *Proceedings of the 12th Python in science conference*, vol. 13, 2013, p. 20.

[43] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.